Multimodal Learning With Transformers: A Survey

Year: 2022 | Citations: 798 | Authors: P. Xu, Xiatian Zhu, D. Clifton

Abstract

Transformer is a promising neural network learner, and has achieved great success in various machine learning tasks. Thanks to the recent prevalence of multimodal applications and Big Data, Transformer-based multimodal learning has become a hot topic in AI research. This paper presents a comprehensive survey of Transformer techniques oriented at multimodal data. The main contents of this survey include: (1) a background of multimodal learning, Transformer ecosystem, and the multimodal Big Data era, (2) a systematic review of Vanilla Transformer, Vision Transformer, and multimodal Transformers, from a geometrically topological perspective, (3) a review of multimodal Transformer applications, via two important paradigms, i.e., for multimodal pretraining and for specific multimodal tasks, (4) a summary of the common challenges and designs shared by the multimodal Transformer models and applications, and (5) a discussion of open problems and potential research directions for the community.